Sistema Computacional de Codificação Automática de Atividades Econômicas

Sistema Computacional de Codificação Automática de Atividades Econômicas

Classificação Automática em CNAE

Apresentação do Projeto SCAE

Participantes no Projeto

Dr. Alberto Fereira De Souza (DI/UFES) (coordenador)

Dr. Elias Oliveira (DCI/UFES)

Dra. **Eliana Zandonade** (DE/UFES)

Dr. Hannu Tapio Ahonen (DI/UFES)

Dr. Felipe Maia Galvão França (COPPE Sistemas/UFRJ)

Dra. Priscila M. V. Lima (COPPE Sistemas/UFRJ)

Dra. Claudine Badue (DI/UFES)

Colaboradores

Dr. Valmir Barbosa (COPPE Sistemas/UFRJ)

Dr. Wagner Meira Jr. (DCC/UFMG)

Estrutura da Apresentação

- Classificar para Quê?
- O Problema e Motivação
- O Projeto SCAE
- Técnicas Utilizadas
- Resultados já Alcançados
- Condições para o Sucesso

Classificar para Quê?

Classificar para Quê?

Os bebês classificam as pessoas que os cercam...

Classificar para Quê?

Os bebês classificam as pessoas que os cercam...

Nós classificamos as entidades com as quais nos relacionamos na natureza...

Classificar para Quê?

Os bebês classificam as pessoas que os cercam...

Nós classificamos as entidades com as quais nos relacionamos na natureza...

Classificamos para melhor racionalizarmos o mundo...
Classificamos para melhor recuperarmos o **conhecimento**...

O Problema e Motivação

Nosso problema é: *ler* o **objeto social** e classificá-lo em uma ou mais das **sub-classes** da tabela CNAE.

O Problema e Motivação

Nosso problema é: *ler* o **objeto social** e classificá-lo em uma ou mais das **sub-classes** da tabela CNAE.

COMÉRCIO VAREJISTA DE EQUIPAMENTOS DE SEGURANÇA PARA RESIDÊNCIA

O Problema e Motivação

Nosso problema é: *ler* o **objeto social** e classificá-lo em uma ou mais das **sub-classes** da tabela CNAE.

ATENDIMENTO TELEFÔNICO NA PRÓPRIA SEDE DA EMPRESA, EFETUANDO O REGISTRO DE SINISTROS DOS SEGURADOS DE TERCEIROS

O Problema e Motivação

Nosso problema é: *ler* o **objeto social** e classificá-lo em uma ou mais das **sub-classes** da tabela CNAE.

PROPORCIONAR PLANOS DE BENEFÍCIOS ASSISTENCIAIS ATRAVÉS DE ESTIPULAÇÃO DE SEGUROS COLETIVOS, CONTRAT OS COM EMPRESAS PRESTADORAS DE SERVIÇOS DE ASSISTÊNCIA MÉDICA, ODONTOLÓGICA E OUTROS, MEDIANTE CONVÊN IOS DE PRESTAÇÃO DE SERVIÇOS; ASSISTÊNCIA FINANCEIRA DE EMERGÊNCIA CONVENIADA COM EMPRESAS ESPECIALIZADAS; ASSISTÊNCIA ESPIRITUAL AOS SEUS ASSOCIADOS, ASSIM COM ENCAMINHAMENTO E ACOMPANHAMENTO A CENTROS DE RECUPERAÇÃO ESPECIAL A DEPENDENTES QUÍMICOS E ALCÓLATRAS.

O Problema e Motivação

Nosso problema é: *ler* o **objeto social** e classificá-lo em uma ou mais das **sub-classes** da tabela CNAE.

A efetiva **leitura**, **análise** e a **interpretação** do conteúdo dos documentos tornou-se um processo extremamente caro, quando feito manualmente.

Resolver esse problema . . .

- Contribui com a desburocratização de processos;
- Facilita o monitoramento *online* dos setores econômicos;
- Racionaliza e precisa a fiscalização.

Seção	Divisões	Descrição da Seção
A	01...03	AGRICULTURA, PECUÁRIA, PRODUÇÃO FLORESTAL, PESCA E AQÜICULTURA
B	05...09	INDÚSTRIAS EXTRATIVAS
C	10...33	INDÚSTRIAS DE TRANSFORMAÇÃO
D	35...35	ELETRICIDADE E GÁS
E	36...39	ÁGUA, ESGOTO, ATIVIDADES DE GESTÃO DE RESÍDUOS E DESCONTAMINAÇÃO
F	41...43	CONSTRUÇÃO
G	45...47	COMÉRCIO; REPARAÇÃO DE VEÍCULOS AUTOMOTORES E MOTOCICLETAS
H	49...53	TRANSPORTE, ARMAZENAGEM E CORREIO
I	55...56	ALOJAMENTO E ALIMENTAÇÃO
J	58...63	INFORMAÇÃO E COMUNICAÇÃO
K	64...66	ATIVIDADES FINANCEIRAS, DE SEGUROS E SERVIÇOS RELACIONADOS
		⋱
P	85...85	EDUCAÇÃO
Q	86...88	SAÚDE HUMANA E SERVIÇOS SOCIAIS
R	90...93	ARTES, CULTURA, ESPORTE E RECREAÇÃO
S	94...96	OUTRAS ATIVIDADES DE SERVIÇOS
T	97...97	SERVIÇOS DOMÉSTICOS

CNAE

Hierarquia

Seção:	B	PESCA
Divisão:	05	PESCA, AQÜICULTURA E SERVIÇOS RELACIONADOS
Grupo:	051	PESCA, AQÜICULTURA E SERVIÇOS RELACIONADOS
Classe:	0511-8	PESCA E SERVIÇOS RELACIONADOS
Subclasse	0511-8/01	**PESCA DE PEIXES**

Notas Explicativas:

Esta Subclasse compreende:

– A pesca de peixes em águas marítimas e em águas continentais

Esta Subclasse compreende também:

– A preparação e conservação do peixe no próprio barco

Esta Subclasse não compreende:

– A captura de crustáceos e moluscos (0511-8/02)

– A preparação do peixe (frigorificado, congelado, salgado, seco) e a fabricação de conservas de peixe em estabelecimentos fabris, inclusive em barcos-fabrica (1514-8/00)

– A preparação de qualquer tipo de farinha de peixe (1514-8/00)

– A criação e cultivo de peixes (0512-6/01)

CNAE

Hierarquia

Seção:	B	PESCA
Divisão:	05	PESCA, AQÜICULTURA E SERVIÇOS RELACIONADOS
Grupo:	051	PESCA, AQÜICULTURA E SERVIÇOS RELACIONADOS
Classe:	0511-8	PESCA E SERVIÇOS RELACIONADOS
Subclasse	0511-8/02	**PESCA DE CRUSTÁCEOS E MOLUSCOS**

Notas Explicativas:

Esta Subclasse compreende:

– A captura de crustáceos e moluscos em águas marítimas e em águas continentais

Esta Subclasse não compreende:

– A preparação de crustáceos e moluscos (frigorificado, congelado, salgado, seco) e a fabricação de conservas de crustáceos e moluscos em estabelecimentos fabris, inclusive em barcos-fábrica (1514-8/00)

Técnicas Alternativas

- Booleano – Termos *vs.* Termos (Google) ;

Técnicas Alternativas

- Booleano – Termos *vs.* Termos (Google) ;
- Estatísticas;

SCAE

Técnicas Alternativas

- Booleano – Termos *vs.* Termos (Google) ;
- Estatísticas;
- **Vetoriais**;
  - Vetoriais Simples;
  - Vetoriais Generalizado;
  - Redes Neurais;
  - Outras.

Entendendo o *Modelo Vetorial*

Nós seres humanos "pensamos", as máquinas "fazem contas"...

Entendendo o *Modelo Vetorial*

Nós seres humanos "pensamos", as máquinas "fazem contas"...
Precisamos transformar o processo de classificação em um processo de contagem...

Entendendo o *Modelo Vetorial*

Nós seres humanos "pensamos", as máquinas "fazem contas"...
Precisamos transformar o processo de classificação em um processo de contagem...

Vamos supor que tenhamos uma base de dados $D = \{d_1, d_2, \ldots, d_j, \ldots, d_n\}$ e queiramos saber quão similar q (um outro documento) é de um ou mais documentos em D.

Entendendo o *Modelo Vetorial*

Nós seres humanos "pensamos", as máquinas "fazem contas"...
Precisamos transformar o processo de classificação em um processo de contagem...

Vamos supor que tenhamos uma base de dados $D = \{d_1, d_2, \ldots, d_j, \ldots, d_n\}$ e queiramos saber quão similar q (um outro documento) é de um ou mais documentos em D.

$$sim(d_j, q) = \frac{\mathbf{d_j} \bullet \mathbf{q}}{|\,\mathbf{d_j}\,| \times |\,\mathbf{q}\,|}$$

Entendendo o *Modelo Vetorial*

Nós seres humanos "pensamos", as máquinas "fazem contas"... Precisamos transformar o processo de classificação em um processo de contagem...

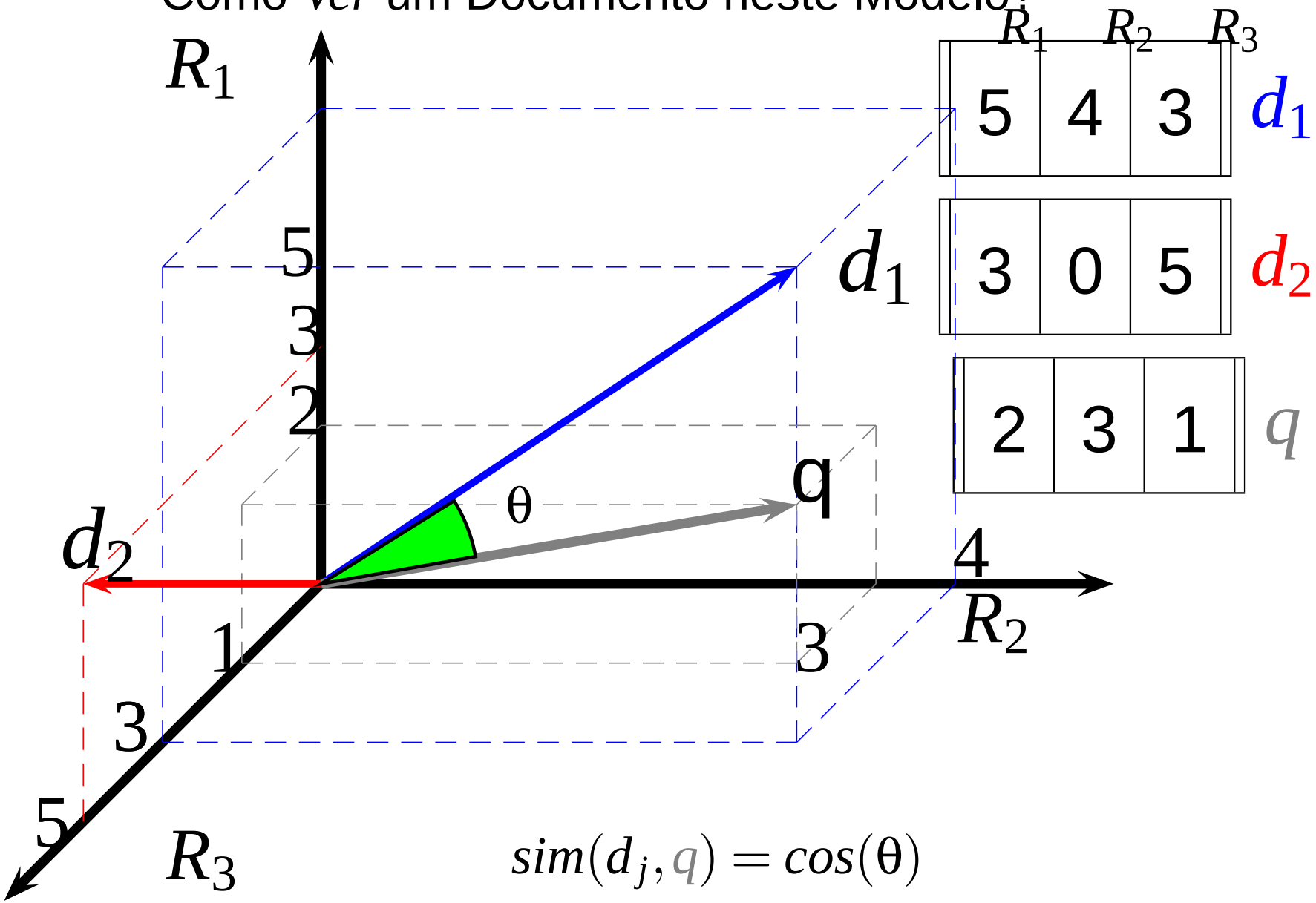
Vamos supor que tenhamos uma base de dados $D = \{d_1, d_2, \ldots, d_j, \ldots, d_n\}$ e queiramos saber quão similar q (um outro documento) é de um ou mais documentos em D.

$$sim(d_j, q) = \frac{\mathbf{d_j} \bullet \mathbf{q}}{|\,\mathbf{d_j}\,| \times |\,\mathbf{q}\,|}$$

$$= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}} = cos(\theta)$$

Entendendo o *Modelo Vetorial*

Como *Ver* um Documento neste Modelo?



$$sim(d_j, q) = cos(\theta)$$

Entendendo o *Modelo Vetorial*

Como Calcular os $w_{i,j}$s

Considere que você tenha $\mid D \mid = N$ documentos na sua base de dados $D = \{d_1, d_2, \ldots, d_j, \ldots, d_n\}$ e que n_i seja o número de documentos em que o termo k_i aparece.

Entendendo o *Modelo Vetorial*

Como Calcular os $w_{i,j}$s

Considere que você tenha $\mid D \mid = N$ documentos na sua base de dados $D = \{d_1, d_2, \ldots, d_j, \ldots, d_n\}$ e que n_i seja o número de documentos em que o termo k_i aparece.

A $freq_{i,j}$ será a frequência do termo k_i no documento d_j. Portanto, a normalização $f_{i,j}$ desta frequência do termo k_i, no documento d_j, é dada por:

$$f_{i,j} = \frac{freq_{i,j}}{max_l \ freq_{l,j}}$$

Entendendo o *Modelo Vetorial*

Como Calcular os $w_{i,j}$s

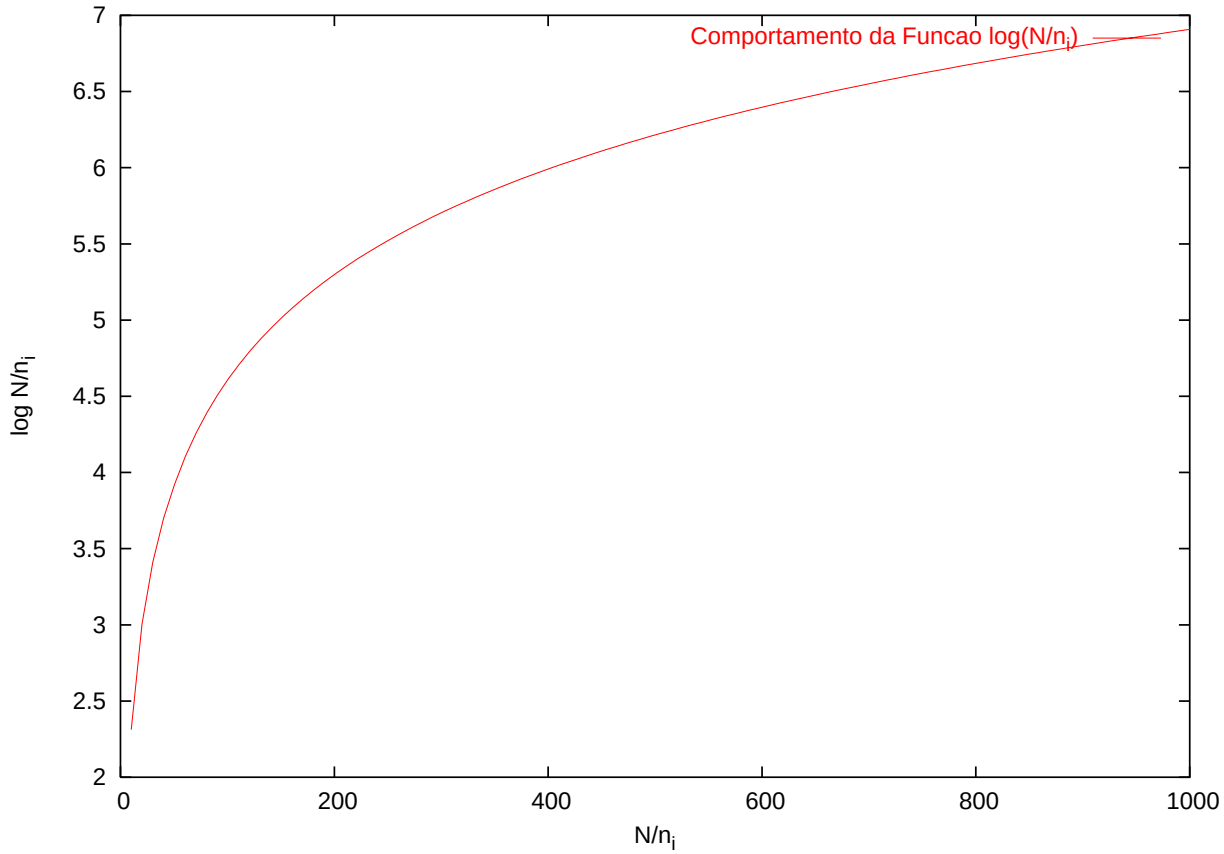
Precisamos também calcular qual a influência deste termo k_i em toda a base de dados. Usaremos para isso o $id\,f_i$ (*inverse document frequency*). Com esta função queremos tornar sensível o fato de que se um termo aparece em todos os documentos, esta função assumirá valor zero.

$$id\,f_i = log\frac{N}{n_i}$$

Entendendo o *Modelo Vetorial*

Como Calcular os $w_{i,j}$s

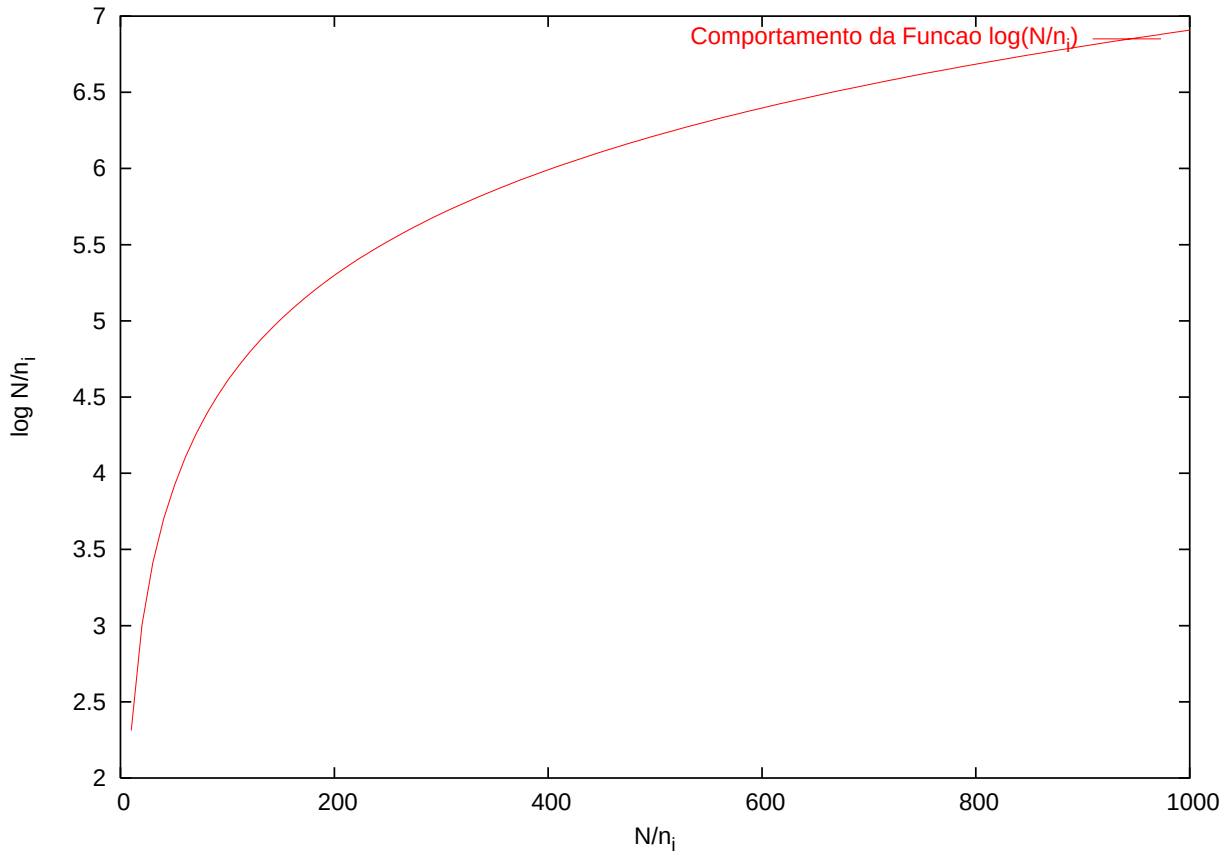
$$id\,f_i = log\frac{N}{n_i}$$



Entendendo o *Modelo Vetorial*

Como Calcular os $w_{i,j}$s

$$w_{i,j} = f_{i,j} \times id\, f_i$$



O projeto SCAE

O objetivo principal deste projeto de pesquisa é desenvolver ou adaptar algoritmos e heurísticas que viabilizem a implementação de um **Sistema Computacional para a Codificação Automática de Atividades Econômicas** (**SCAE**) e comparar o desempenho deste sistema com o de codificadores humanos.

Assim, decidimos atacar duas frentes de trabalho:

1. Desenvolvimento algoritmos e heurísticas para classificação automática;
2. Desenvolvimento de um sistema de apoio para treinamento e **avaliação** de classificadores humanos.

SCAE

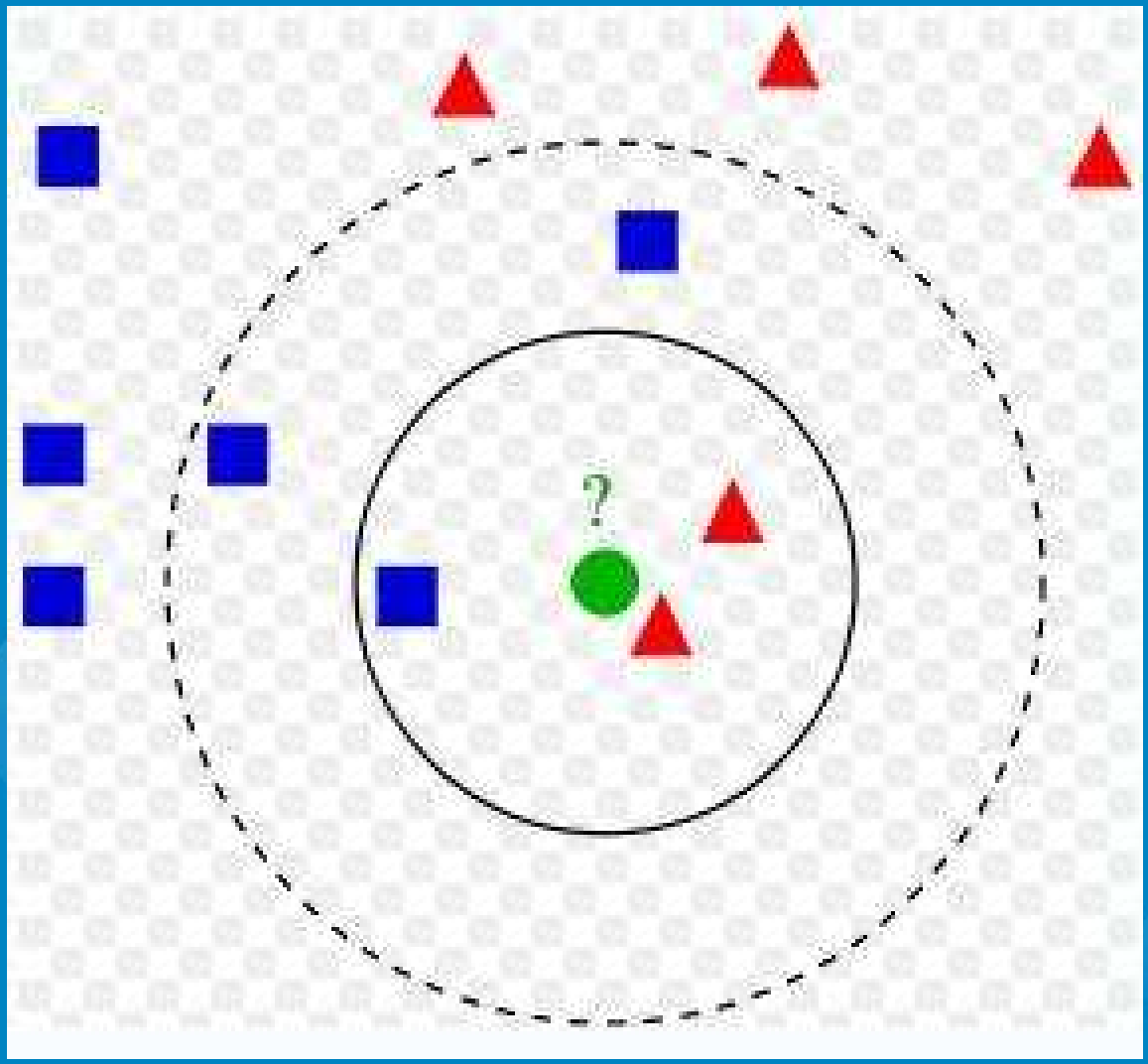
Técnicas Utilizadas

Cada um dos pesquisadores relacionados anteriormente é responsável por uma técnica de classificação automática. Portanto, temos mais de 5 técnicas sendo trabalhadas no momento.

1. kNN – k Nearest Neighbor Algorithm;
2. ML-kNN – Multi-Label kNN;
3. PNN – Probabilistic Neural Network;
4. VG-RAM – Virtual Generalizing RAM networks;
5. Bayesian network;
6. Bootstrap;
7. ...

SCAE

Como Funciona o kNN?



Resultados já Obtidos

CIARELLI, P. M. et al. Uma Biblioteca Digital de Objetos Sociais de Empresas e a Classificação Automática Nacional de Atividades Econômicas. In: *V Simpósio Internacional de Bibliotecas Digitais*. São Paulo: [s.n.], 2007.

OLIVEIRA, E. et al. Comparison Between a kNN based Approach and a PNN Algorithm for a Multi-Label Classification Problem. In: *8th International Conference on Intelligent Systems Design and Applications*. Rio de Janeiro: IEEE Computer Society, 2008. p. 628–633.

De Souza, A. F. et al. Automated Free Text Classification of Economic Activities using VG-RAM Weightless Neural Networks. In: *7th International Conference on Intelligent Systems Design and Applications*. Rio de Janeiro: [s.n.], 2007.

OLIVEIRA, E. et al. Intelligent Classification of Economic Activities from Free Text Descriptions. In: *5º Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*. Rio de Janeiro: [s.n.], 2007.

CIARELLI, P. M.; LIMA, F. O.; OLIVEIRA, E. Using a Genetic Algorithm for Configuring a Set of Probabilistic Neural Networks. In: *XXXIX SOBRAPO*. Fortaleza, Ceará: [s.n.], 2007.

CIARELLI, P. M.; OLIVEIRA, E. Using a Probabilistic Neural Network for a Large Multi-label Problem. In: *10th Brazilian Symposium on Artificial Neural Networks (SBRN)*. Salvador, Bahia: [s.n.], 2008.

Condições para o Sucesso

Os algoritmos que utilizamos são baseados em aprendizado. Portanto, é preciso que os apresentemos **muitos bons** exemplos para que os mesmos possam aprender **corretamente**.

SCAE

Base de Dados: Problema Estatístico

Definir uma base de dados **representativa** que tenha como missão:

1. aferir a capacidade de classificação dos classificadores manuais; **B1** e **B3**
2. estudar os modelos algoritmicos quanto a sua capacidade de resolver o problema proposto no projeto, considerando as particularidades; **B2** e **B4**
3. calibrar os modelos automatizados propostos (baseados em aprendizado); **B4**
4. aferir a capacidade de classificação dos modelos propostos;
5. comparar estatisticamente estes modelos.

SCAE

Base de Dados: B1

Dados da central de dúvidas do IBGE (texto da atividade principal, perguntas do IBGE e o código atribuído pelos especialistas).

Meta: **360** e-mails
Alcançados: **101**

Dificuldade	Área				TOTAL
	serviço	indústria	comércio	agricultura	
Fácil	30	30	30	30	120
Médio	30	30	30	30	120
Difícil	30	30	30	30	120
TOTAL	90	90	90	90	**360**

SCAE

Base de Dados: B2

Dados de objeto social das prefeituras de **Vitória**, **Belo Horizonte** e **Junta Comercial de Sergipe**.

Esses dados consistem nos textos das atividades e códigos atribuídos pelos respectivos órgãos.

Combinação das bases de **Vitória** e **BH** para gerar uma nova base com mais exemplos em determinados códigos.

SCAE

Base de Dados: B3

Dados da pesquisa econômica do IBGE. Em torno de **30 mil** empresas. Dados da atividade principal, perguntas e códigos.

Meta: ter os dados com os códigos

Alcançado:

- dados com códigos: **9.000**
- dados sem o código: 27.246

Base de Dados: B4

Base montada para os experimentos do projeto.

Essa base será coletada por intermédio de um protótipo de entrada de dados com texto livre de atividades, perguntas adicionais e estes **dados serão classificados pelos classificadores humanos**.

Tamanho da Base B4

Proposta 1 Pelo menos 10 repetições de cada código (1183) para todas as capitais (27):

$$10 \times 1183 \times 27 = 319.410$$

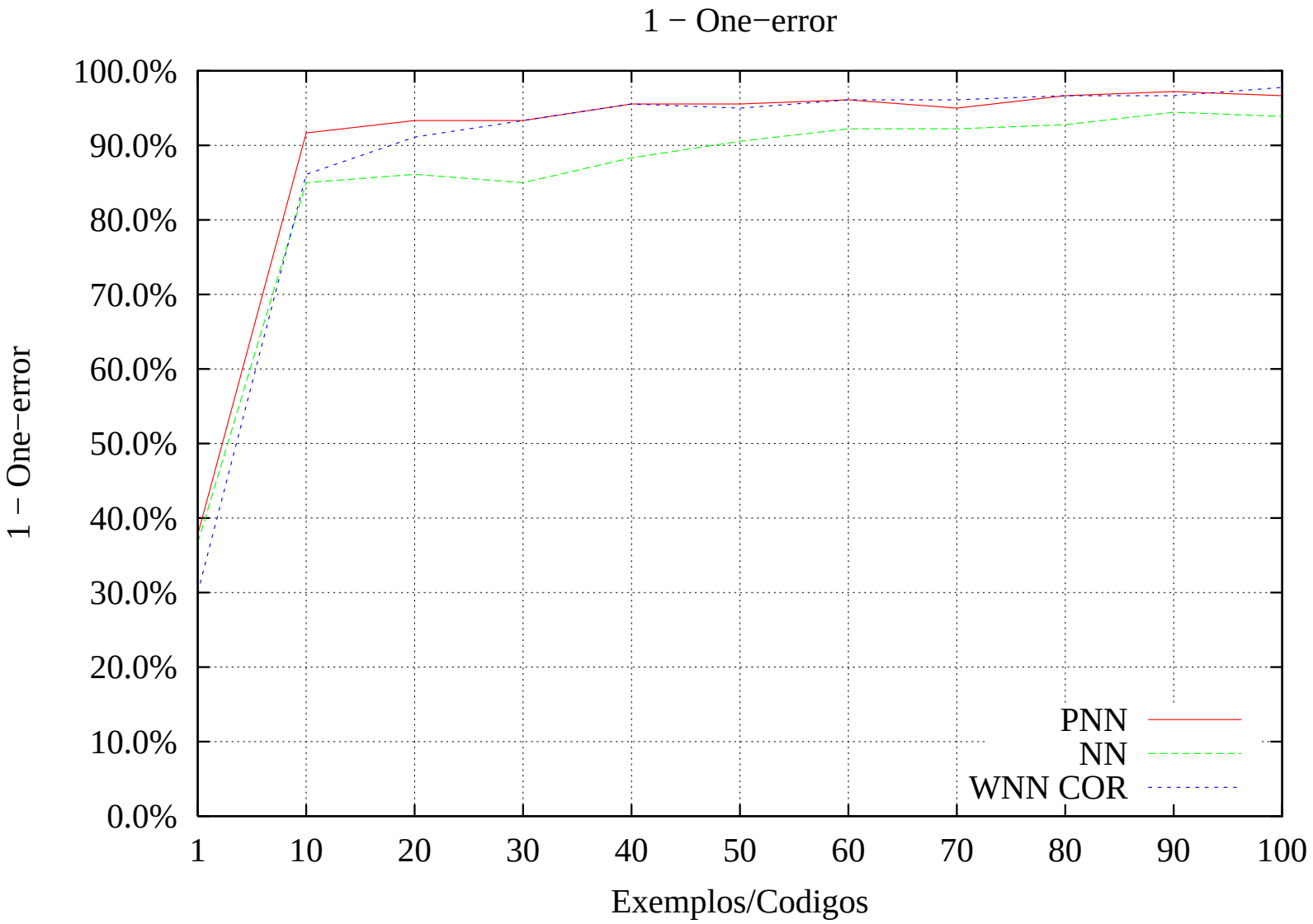
Proposta 2 Pelo menos 30 repetições de cada código (1183) para todas as regiões (5):

$$30 \times 1183 \times 5 = 177.450$$

Proposta 3 Pelo menos 100 repetições de cada código (1183) para todas as regiões (5):

$$100 \times 1183 \times 5 = 591.500$$

Aprendizado *vs.* Número de Exemplos



SCAE

Desafios

1. Obtenção de bons exemplos que, de fato, **representem todas** as atividades econômicas;
2. Melhorar o desempenho dos algoritmos para que possam **aprender dinâmicamente** mais e mais rápido com a inserção de novos conhecimentos;
3. Lidar com grandes bases de conhecimentos: Tera Bytes...
4. Garantir para o usuários, com **comprovações estatísticas**, um bom **nível de qualidade** de sugestão de classificação

OBRIGADO!